Analyse Corpus Scientifique Escarpit

Mélanie Aubry ; Cécile Portal

10/01/2019

Description du dossier

Ce dossier est une analyse de corpus réalisée sur Iramuteq. Ce corpus contient 23 textes de Robert Escarpit classés comme *"Articles et Communications scientifiques"*. C'est un compte-rendu de notre expérience qui présente ce que nous avons fait, ce que nous avons obtenu et notre avis. Vous pourrez retrouver tout notre dossier sur **GitHub** en cliquant ici. Un autre fichier .md différent de GitHub a été créé pour faire marcher Pandoc. Voici ci-dessous les deux commandes **Pandoc** utilisées :

```
C:\Users\Mao>pandoc -s -o test2.pdf analyse_pandoc.md
```

```
C:\Users\Mao>pandoc -s -o test5.pdf --css pandoc.css analyse_pandoc.html
```

Pour convertir en PDF, nous avons été obligées de télécharger Miktex. Le seul souci de la version PDF est le positionnement des images, qui n'est pas comme sur notre markdown ou notre GitHub. En effet, les images se retrouvent décalées, et malgré plusieurs tests avec d'autres convertisseurs de Pandoc, le souci reste le même. Un test rapide en html a été fait pour vérifier le positionnement qui est correct dans cette version. Nous n'avions malheureusement pas exporté les images en .svg, pensant que le PDF irait.

Choix des métadonnées

Variables

Nous avons regroupé tous les textes dans des fichiers .txt (trouvables ici) et encodé ceux-ci de la façon suivante :

```
**** *variableX *variableX.1
texte texte texte texte

**** *variableY *variableY.2
texte texte texte
```

```
A noter que :
**** : introduit chaque texte
*nomvariable : crée une variable
```

De plus nous avons intégré les métadonnées suivantes, en fonction des dates de publication de chacun des textes, de leurs sujets et de leur type : * *date_XXXX* *subject_XXXX* *type_XXXX

Notre choix de variables repose donc sur différentes catégories. En voici un exemple :

*date_XXXX	*subject_XXXX	*type_XXXX
1950	infocom	articles scientifiques
1970	histoire	article
1990	sociologie	discours

Regroupement de textes

Nous avons décidé de procéder à plusieurs analyses en créant des **sous-corpus**. Tout d'abord, nous choisirons les méta-données de langue. Nous séparerons tout d'abord les 19 textes français des 3 textes en anglais et de celui en italien par souci de pertinence (dictionnaires sur Iramuteq) et de compréhension. Nous les comparerons dans une petite conclusion. Nous ne retiendrons que **les formes actives**, car nous pensons que les formes supplémentaires ne sont **pas pertinentes** pour une analyse. Enfin, dans seconde partie, nous nous attarderons sur la totalité des textes en faisant une analyse par période, pour suivre l'évolution des thématiques Robert Escarpit au fil du temps. Après un bref compte-rendu, nous donnerons notre avis sur les travaux réalisés pour ce devoir.

Analyse selon la langue

Analyses statistiques

Ce type d'analyse permet d'avoir plus de **lisibilité** en matière de compréhension et d'analyse du texte. La lemmatisation permet de réduire les verbes à leurs formes infinitives. Nous avons alors obtenu un schéma ainsi que plusieurs tableaux CSV permettant de voir la fréquence des mots dans l'oeuvre ainsi que leurs types (verbe, adjectif. . .).

Analyse des textes en français

Résumé :

- Nombre de textes : 19
- Nombre d'occurences : 88001
- Nombre de formes : 7511

- Nombre d'hapax : 3410 (3.87% des occurences - 45.40% des formes)
- Moyenne d'occurences par texte : 4631.63

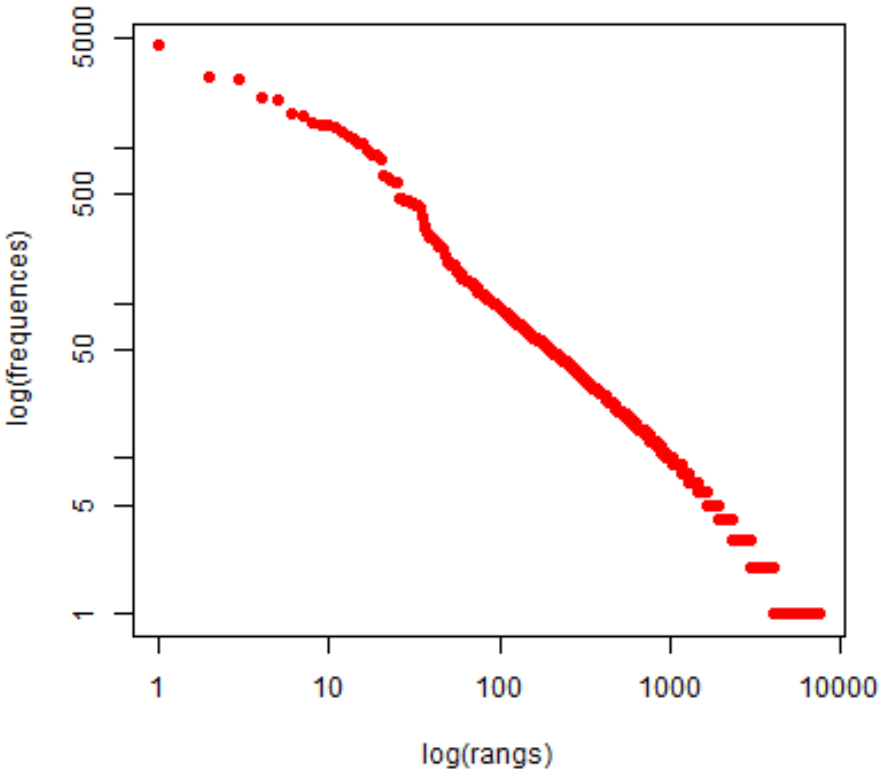

Figure 1: statistiquefr

Analyse des textes en anglais

Résumé :

- Nombre de textes : 3
- Nombre d'occurences : 10207
- Nombre de formes : 2244
- Nombre d'hapax : 1210 (11.85% des occurences - 53.92% des formes)
- Moyenne d'occurences par texte : 3402.33

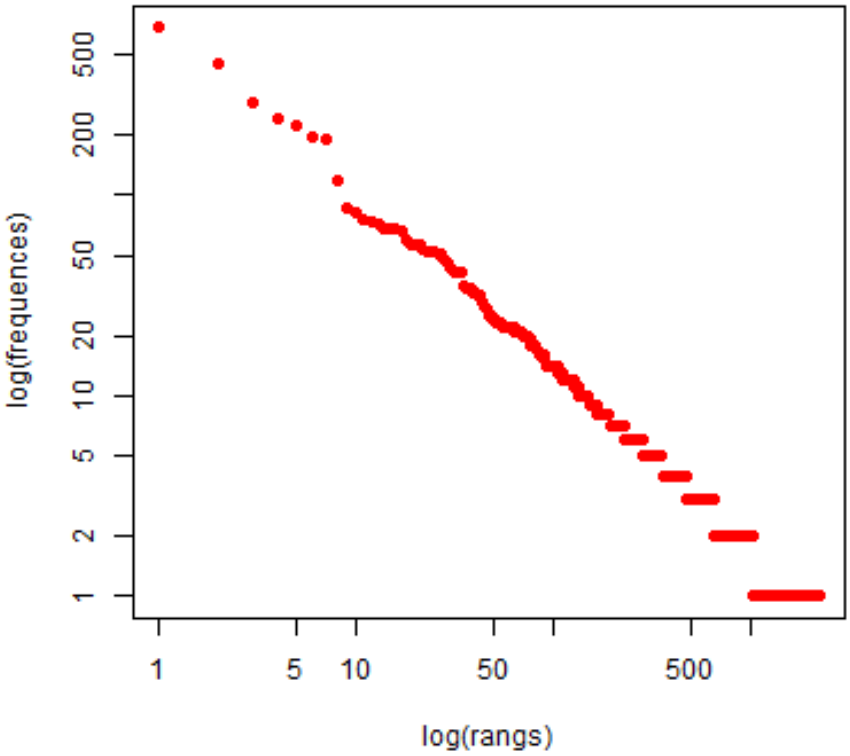

Figure 2: statistiqueeng

Analyse du texte en italien

Résumé :

- Nombre de textes : 1
- Nombre d'occurences : 4321
- Nombre de formes : 1200
- Nombre d'hapax : 705 (16.32% des occurences - 58.75% des formes)
- Moyenne d'occurences par texte : 4321.00

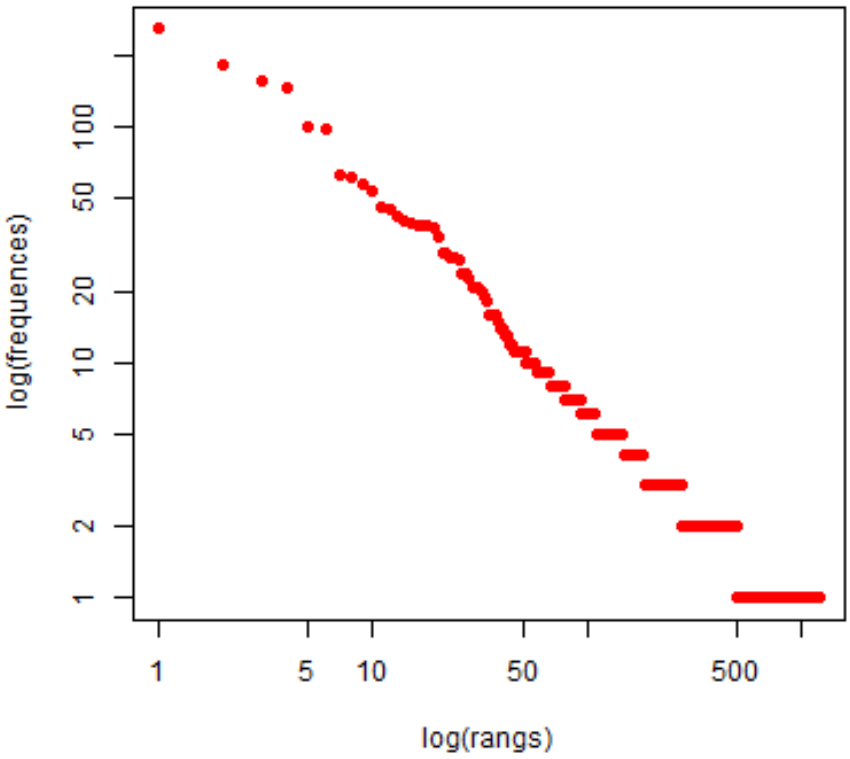

Figure 3: statistiqueita

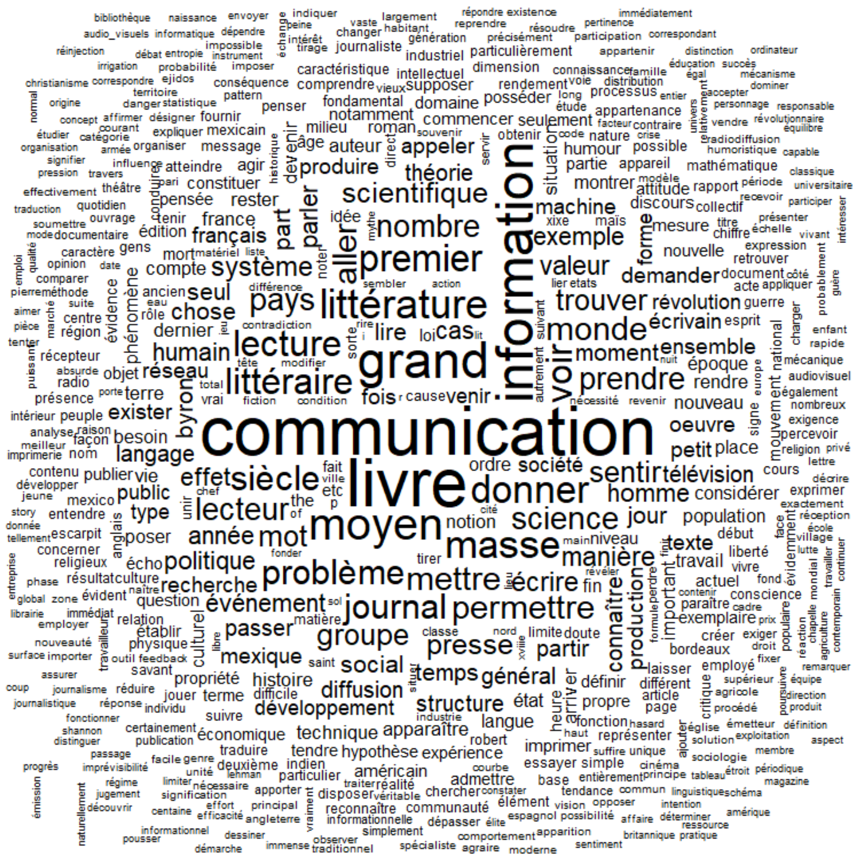

Figure 4: nuagefr

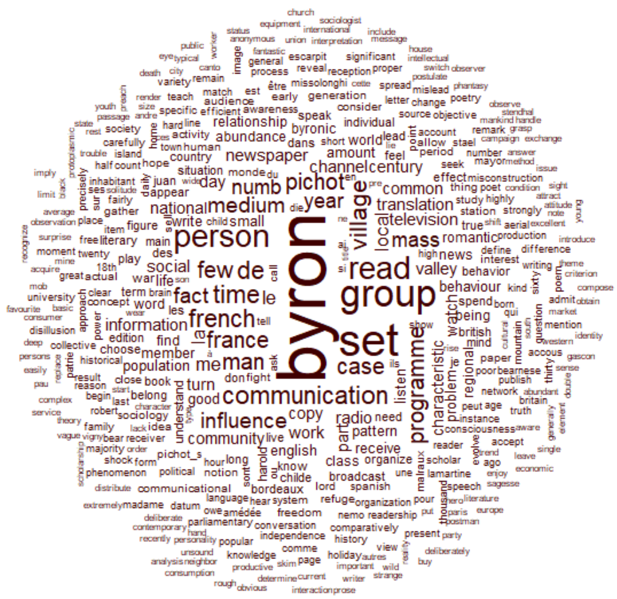

Figure 5: nuageeng

Nuage de mots

Nuage de mots en français

Nuage de mots en anglais

Nuage de mots en italien

Figure 6: nuageita

Analyse méthode Reinert

La classification de Reinert permet de classer les formes dans des **classes de formes** regroupées selon leur indépendance mesurée par un test au Chi². Ces mêmes classes peuvent alors être représentées à l'aide de différents arbres, comme ici avec des dendrogrammes des classes lexicales. Ce diagramme fournit la liste des formes les plus associées pour chaque classe. A noter qu'une forme peut se

retrouver dans **plusieurs** classes différentes. Une classe est un regroupement de segments de texte qui contiennent des formes. Le graphique ci-dessus facilite le repérage des formes et leur degré de dépendance aux classes.

Graphe des classes avec taille des mots proportionnelle à leur fréquence

Français

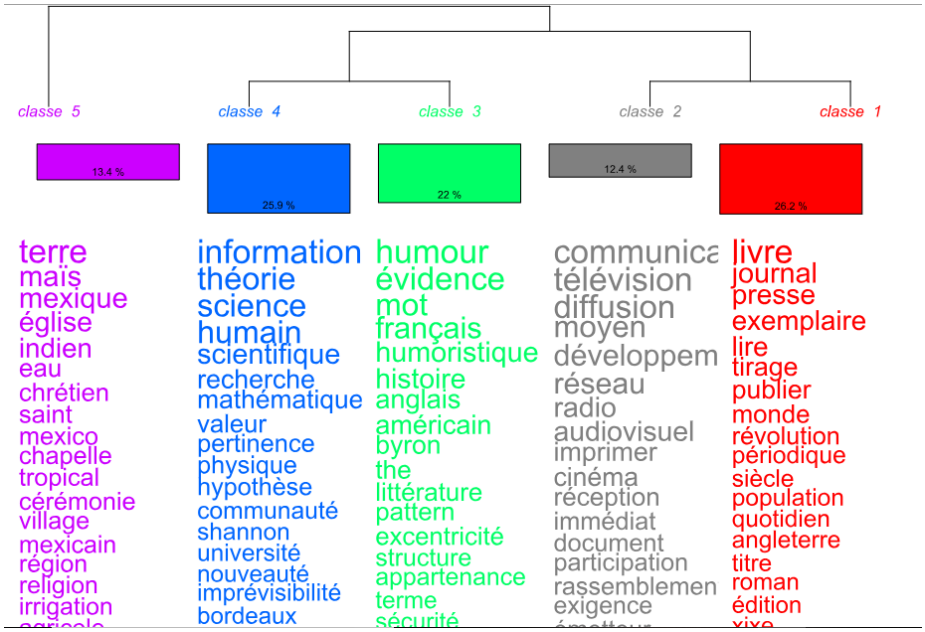

Figure 7: dendrofr

Anglais

Italien

Graphe des classes avec taille des mots proportionnelle au score de chi-2

Avec ce classement par dendrogramme, on peut par la suite trouver une analyse factorielle des correspondances (AFC) reliée au **Chi²**, car le tableau donné par la classification de Reinert utilise des classes dans le tableau lexical. Ce type d'analyse va transformer les données sous forme de graphique a 2 dimensions, montrant la différence entre chaque groupe ou chaque classe de mots dans le but de hiérarchiser les informations des textes. On utilise pour cela plusieurs paramètres comme la fréquence des mots ou encore le type de variables.

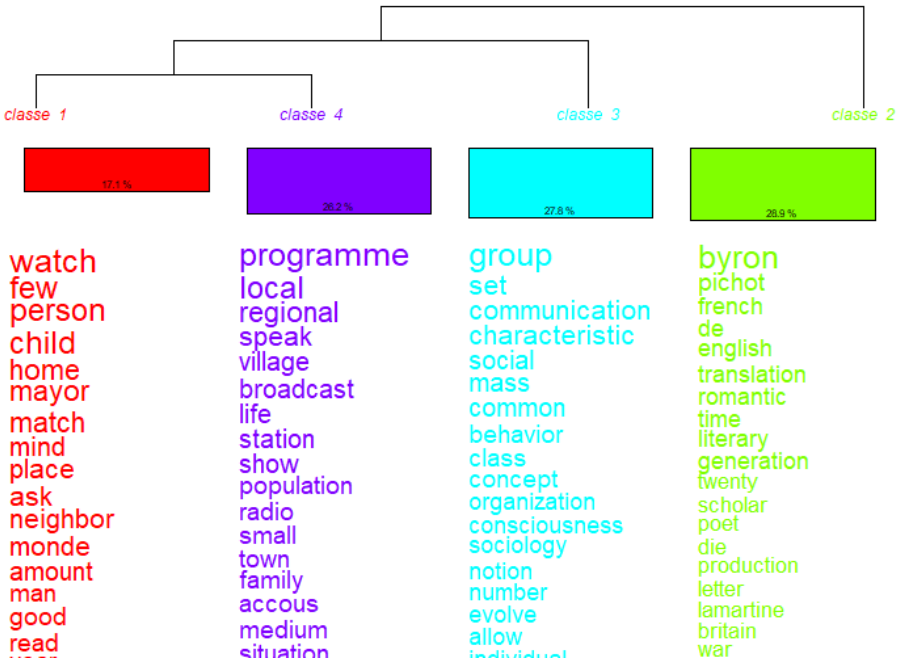

Figure 8: dendroeng

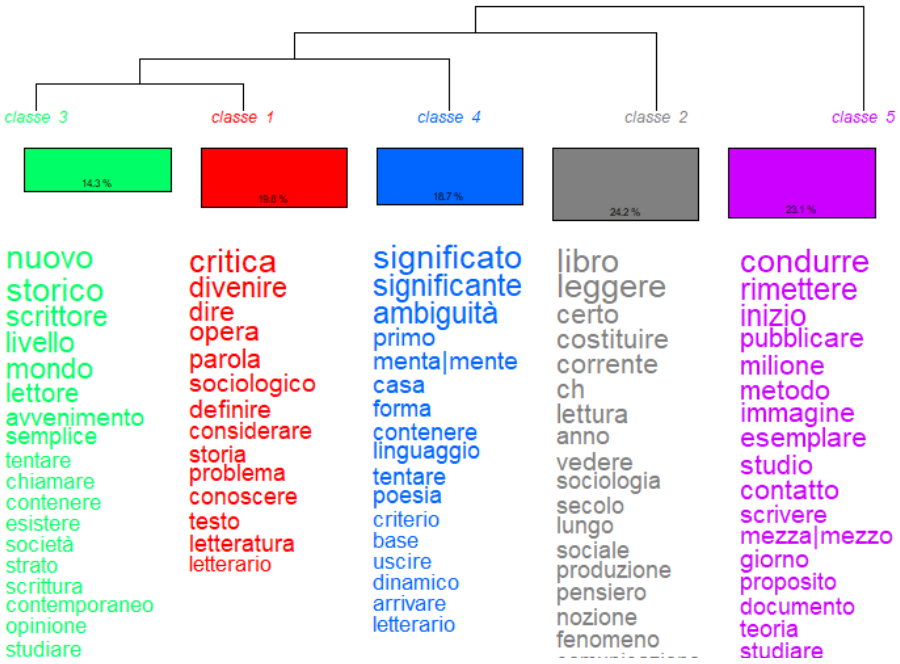

Figure 9: dendroita

Français

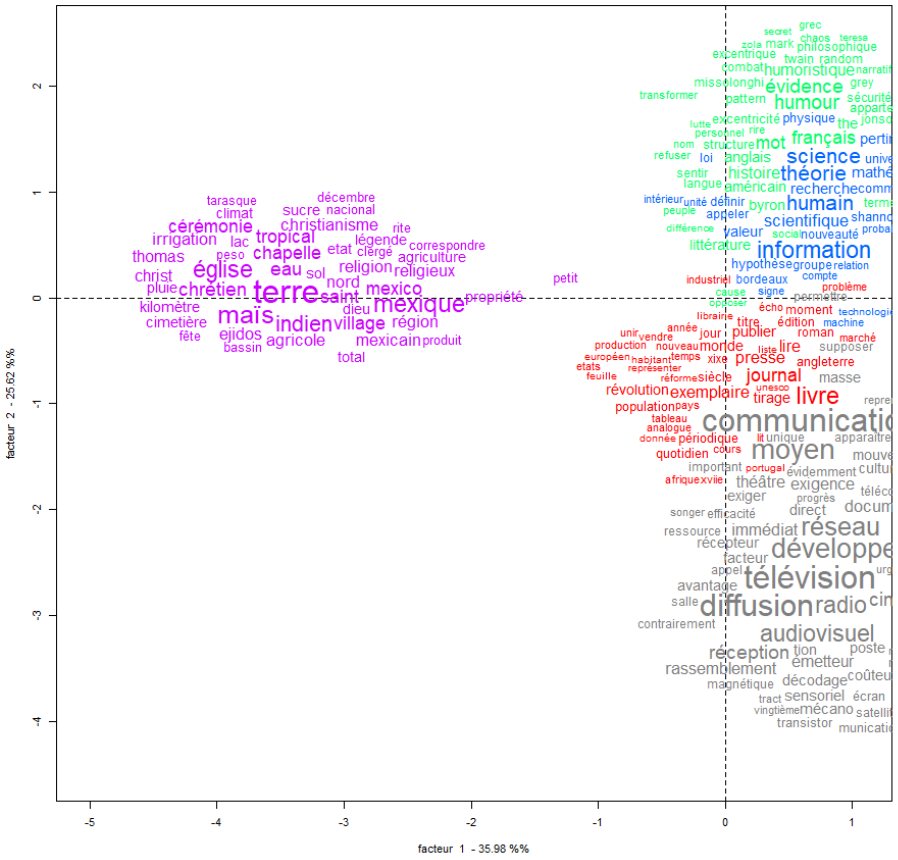

Figure 10: graphfr

Anglais

Italien

Raisonnement

Nous avons tenté de comparer les textes dans différentes langues à l'aide de graphiques. Le point commun entre tous ces textes en anglais, en italien et français sont les thèmes de la **communication**, de la **masse** et du **livre**, bien que les articles divergent. En effet, les textes en anglais s'attardent en particulier sur Byron, tandis que celui en italien traite de la littérature en général. On peut

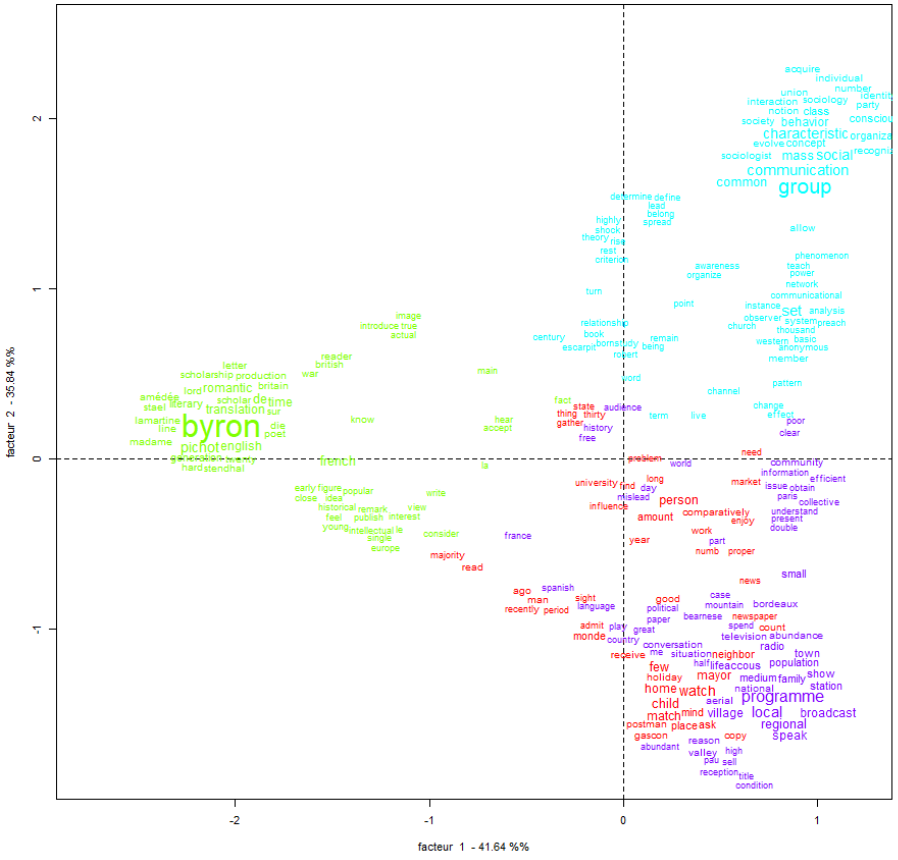

Figure 11: grapheng

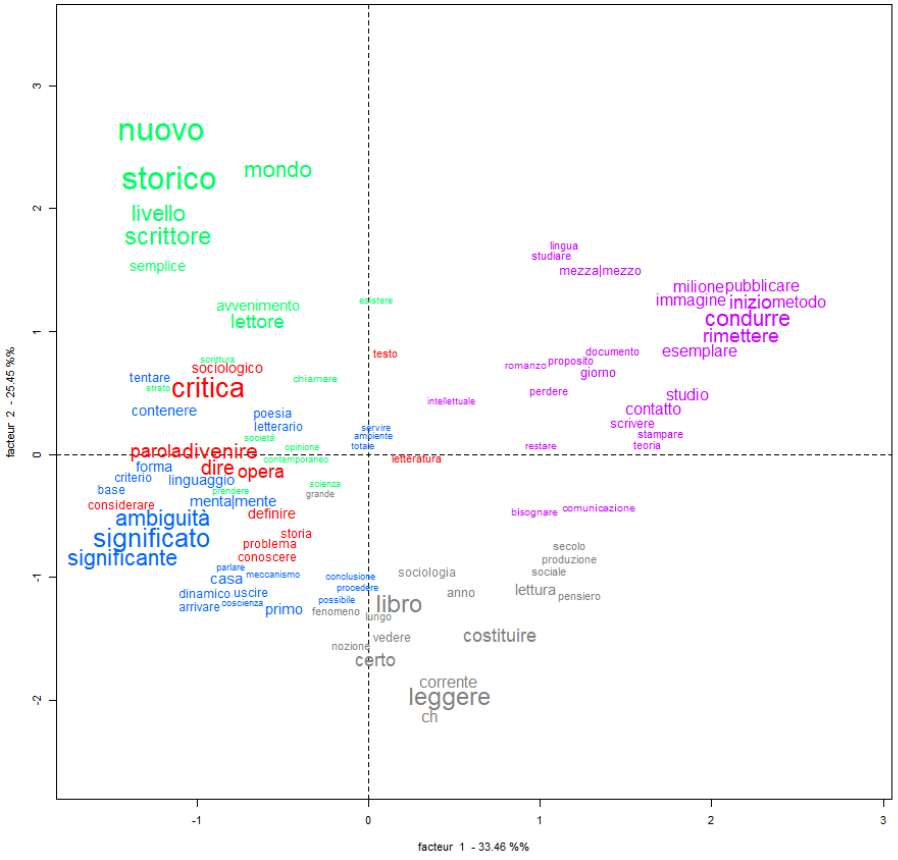

Figure 12: graphita

voir que dans ses textes en français, Escarpit reste focalisé sur les thèmes qui lui sont chers : le **livre**, l'**information** et la **communication**. Les trois différents dendrogrammes possèdent quasiment le même nombre de classes, et on peut observer des similitudes entre le dendrogramme français et l'italien. Quant aux AFC, on peut observer par exemple que le mot **communication** a tendance à se retrouver au même endroit sur les graphiques pour les textes français et italien, un peu moins sur les textes en anglais.

Pour terminer, nous dirons que quelque soit la langue, Robert Escarpit reste fidèle aux thèmes qui le touchent, mais cela ne l'empêche en rien d'avoir écrit de nombreux articles totalement différents dans leur fond.

Analyse Chronologique des articles de Robert Escarpit

1ère période : de 1940 à 1970

Analyse statistique

Résumé :

- Nombre de textes : 6
- Nombre d'occurences : 35769
- Nombre de formes : 4624
- Nombre d'hapax : 2240 (6.26% des occurences - 48.44% des formes)
- Moyenne d'occurences par texte : 5961.50

Nuage de mots

Analyse méthode Reinert

Graphe des classes avec taille des mots proportionnelle à leur fréquence

Graphe des classes avec taille des mots proportionnelle au score de chi-2

2ème période : de 1970 à 1985

Analyse statistique

Résumé :

- Nombre de textes : 14
- Nombre d'occurences : 57364
- Nombre de formes : 7678
- Nombre d'hapax : 3977 (6.93% des occurences - 51.80% des formes)
- Moyenne d'occurences par texte : 4097.43

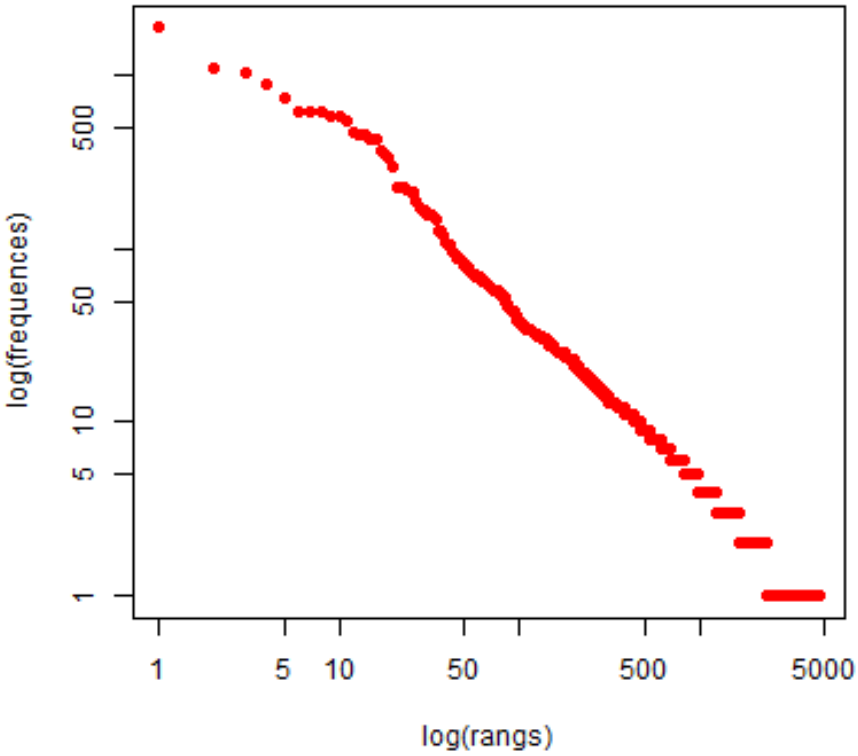

Figure 13: stats40-70

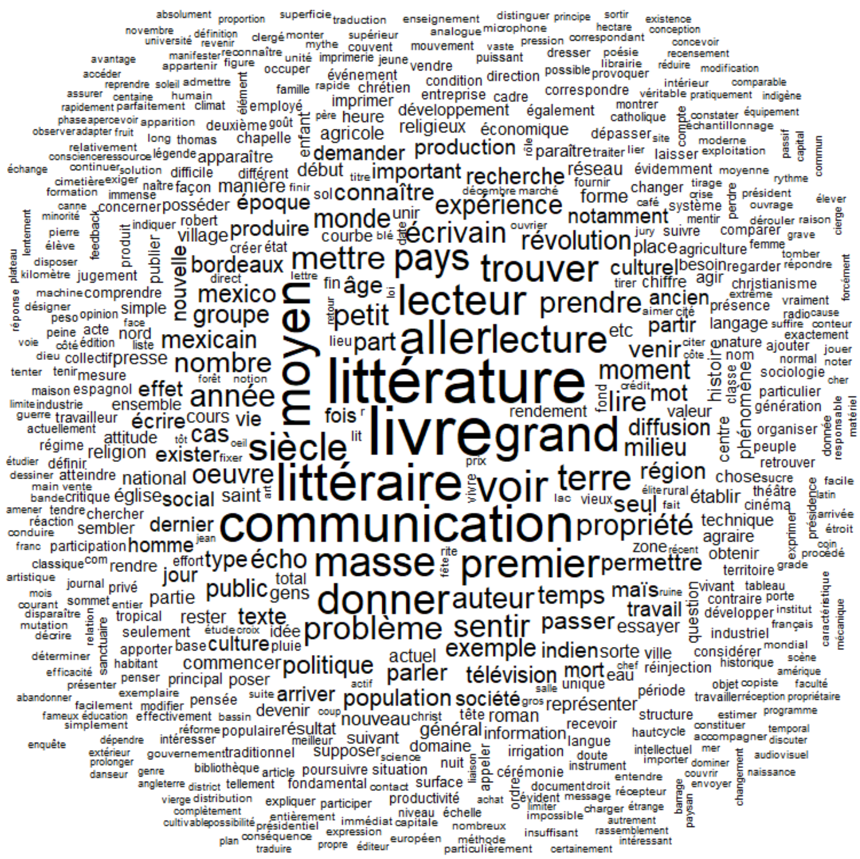

Figure 14: nuage40-70

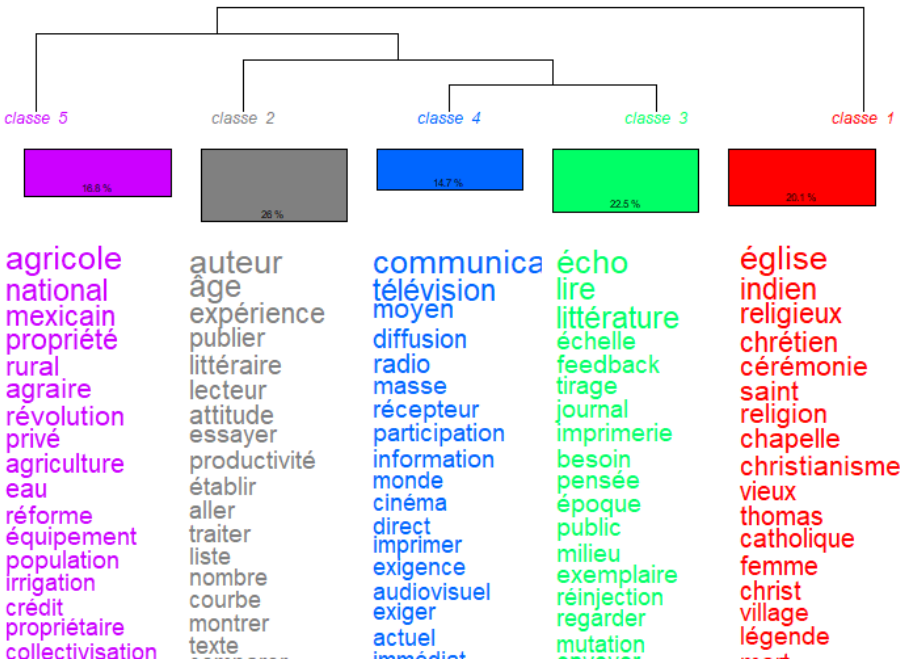

Figure 15: dendro40-70

Nuage de mots

Analyse méthode Reinert

Graphe des classes avec taille des mots proportionnelle à leur fréquence

Graphe des classes avec taille des mots proportionnelle au score de chi-2

Pour ce graphique, nous avons regretté le fait qu'on ne puisse sélectionner qu'un seul dictionnaire, français en l'occurence.

3ème période : à partir de 1985

Analyse statistique

Résumé :

- Nombre de textes : 3
- Nombre d'occurences : 13477
- Nombre de formes : 2345
- Nombre d'hapax : 1240(9.20% des occurences - 52.88% des formes)
- Moyenne d'occurences par texte : 4492.33

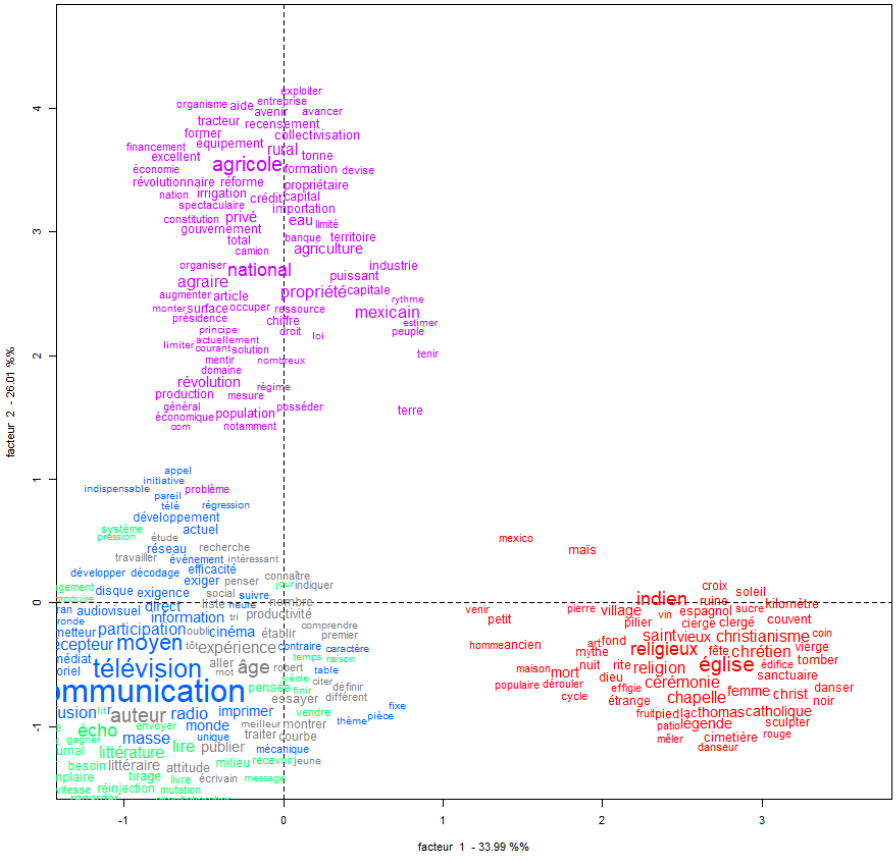

Figure 16: chi2-40-70

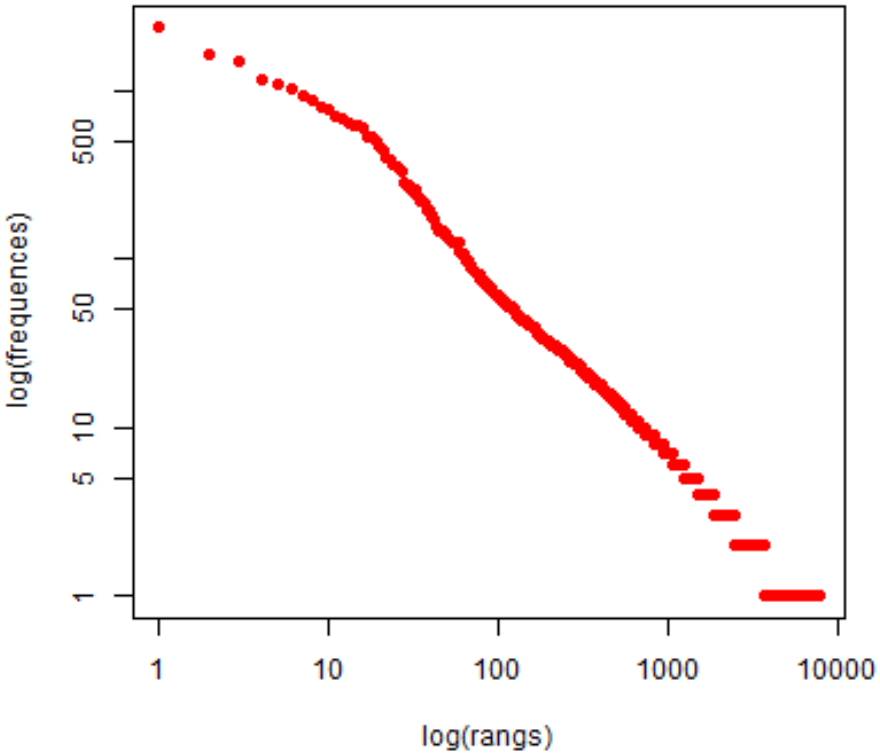

Figure 17: stats70-85

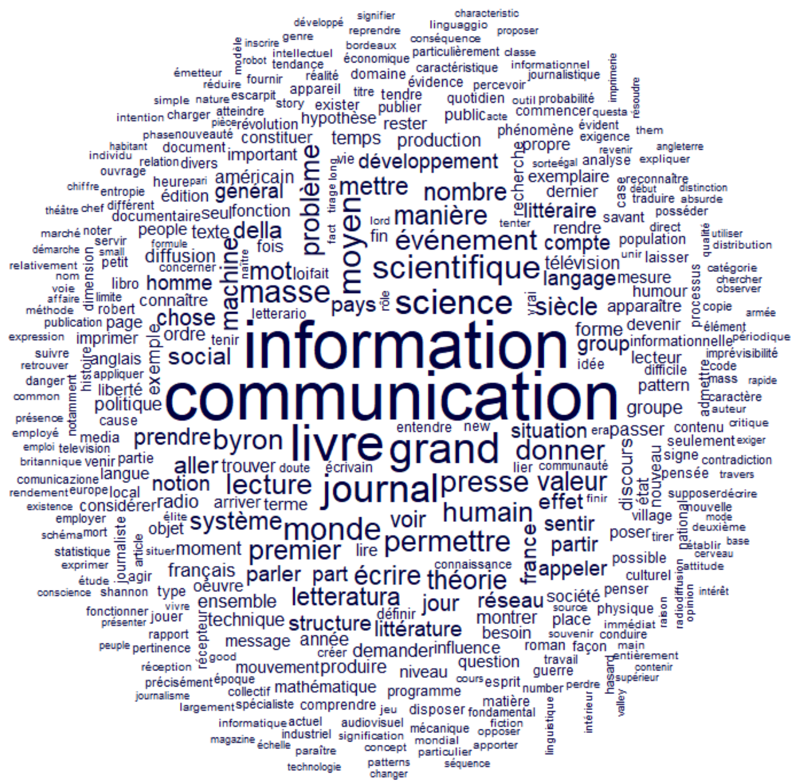

Figure 18: nuage70-85

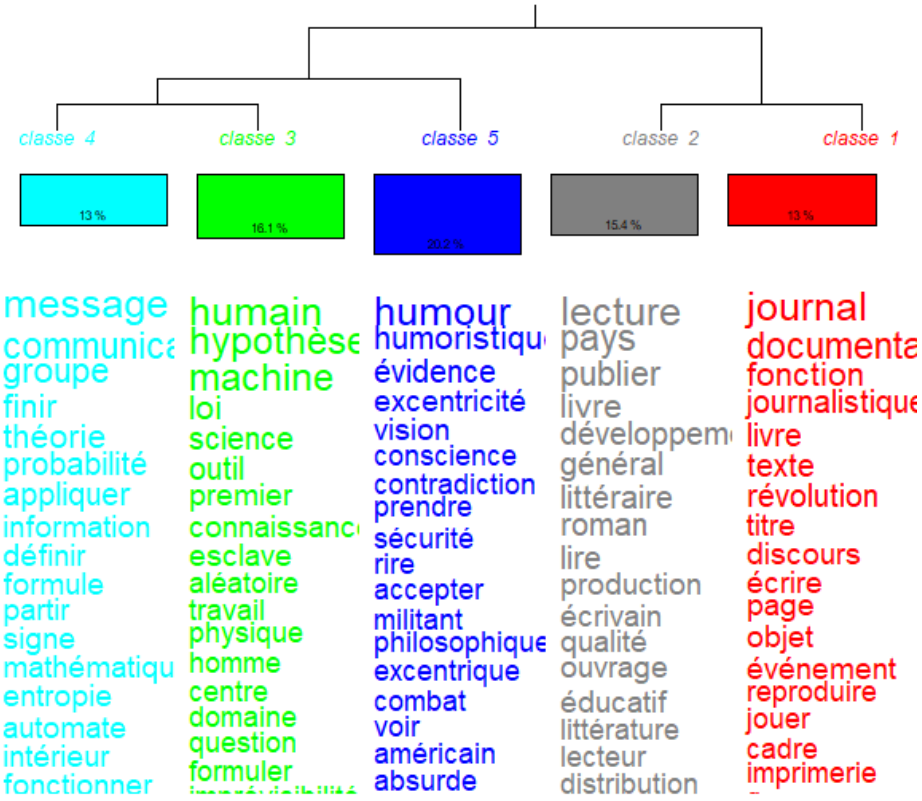

Figure 19: dendro70-85

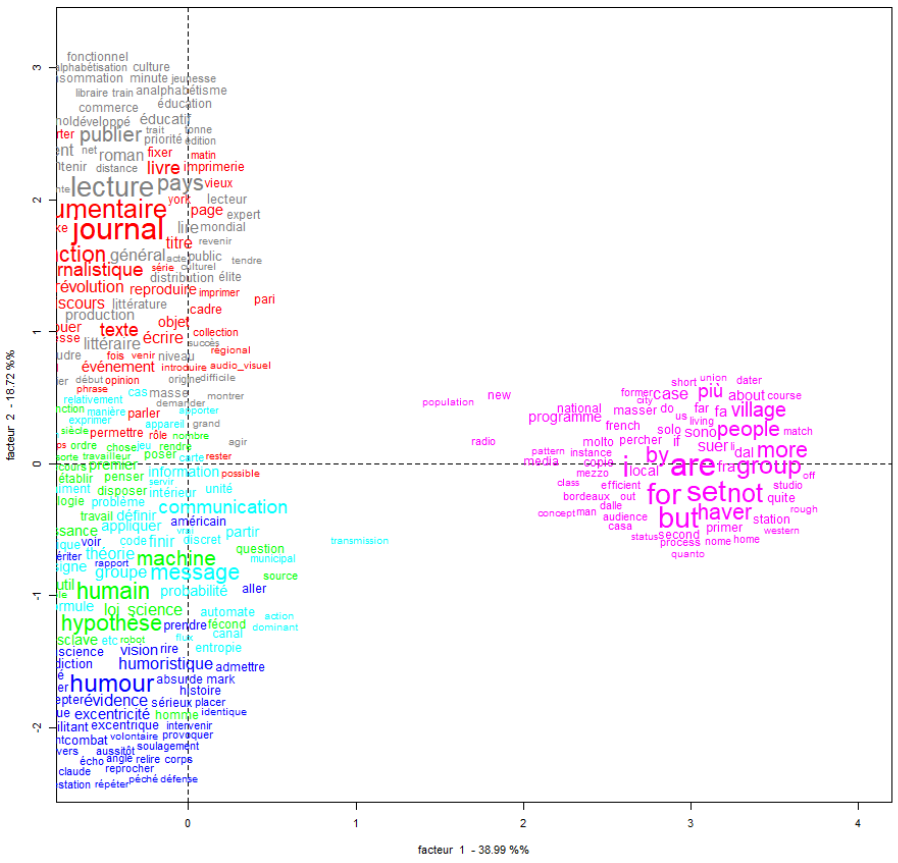

Figure 20: chi2-70-85

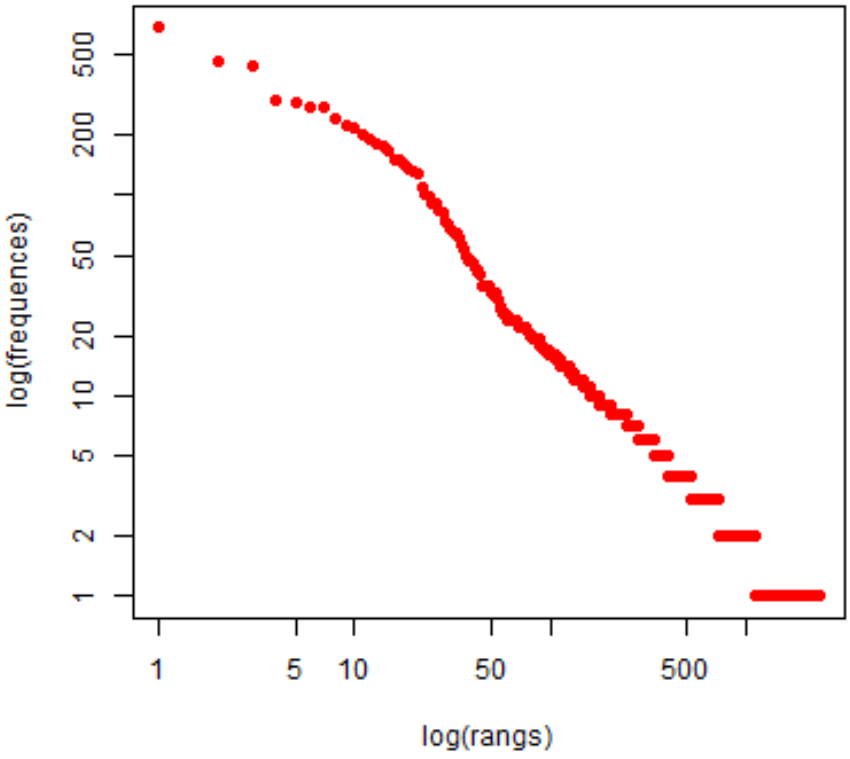

Figure 21: stats85

Nuage de mots

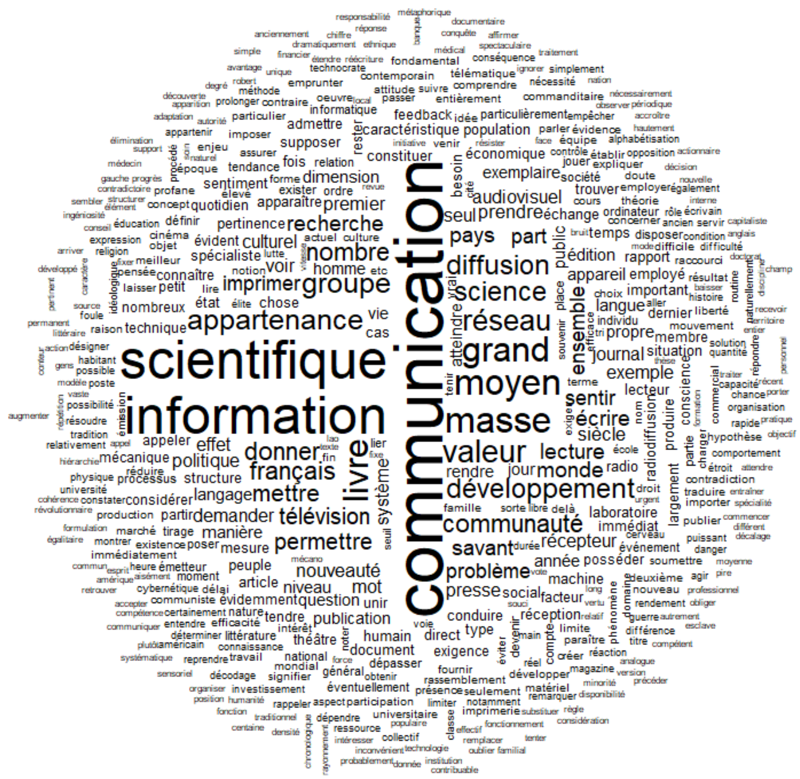

Figure 22: nuage85

Analyse méthode Reinert

Graphe des classes avec taille des mots proportionnelle à leur fréquence

Graphe des classes avec taille des mots proportionnelle au score de chi-2

Raisonnement

Nous avons sélectionné 3 périodes pour cette analyse : de 1940 à 1970, de 1970 à 1985, et de 1985 à la fin. Grâce aux différents graphiques que nous avons

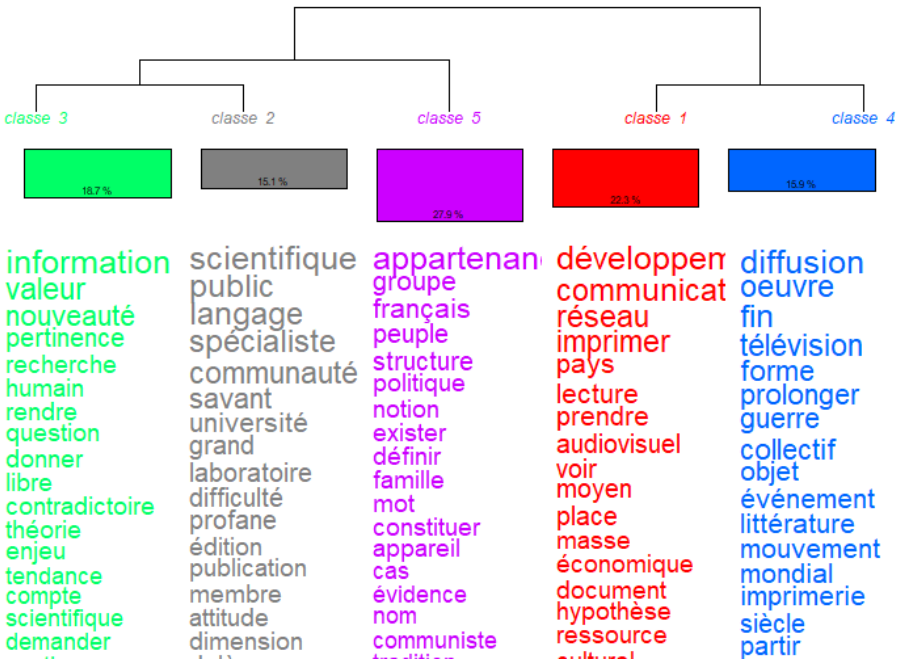

Figure 23: dendro85

pu créer, nous pouvons observers plusieurs choses sur les écrits scientifiques de Robert Escarpit. Tout d'abord, un thème commun à chaque période est celui de la **communication**. En effet, Robert Escarpit a dévoué sa vie à analyser ce phénomène. Ensuite, nous pouvons nous apercevoir que durant la 1ère période (1940-1970), R. Escarpit a préféré rester dans les thèmes de la *littérature*, du *livre* et de la *masse*. Durant la 2ème période (1970-1985), il va se pencher sur l'**information**, son rôle dans la communication, mais garde aussi un intérêt pour le **livre**. Il commence également à y voir de la **science**, et à partir de la 3ème période, le côté scientifique se concrétise et voit son intérêt porté au même niveau que celui de la communication. Durant cette dernière, il va s'intéresser à la notion de *développement*, de *communauté* et d'*appartenance*. Tout au long de son écriture, il garde des notions, comme celle de la **masse** (même si moins dans la suite de sa vie) et de **moyen**.

Conclusion

Cela a parfois été un peu compliqué de procéder à l'analyse, car certains textes ont été très mal corrigés et nous ont quelque peu ralenties. Mais tous les travaux que nous avons pu effectuer sur Iramuteq nous ont montré que l'aide logicielle à l'analyse est très **prometteuse** pour l'avenir. En plus d'apporter une analyse pertinente des oeuvres selon plus axes, l'analyse logicielle permet d'exploiter

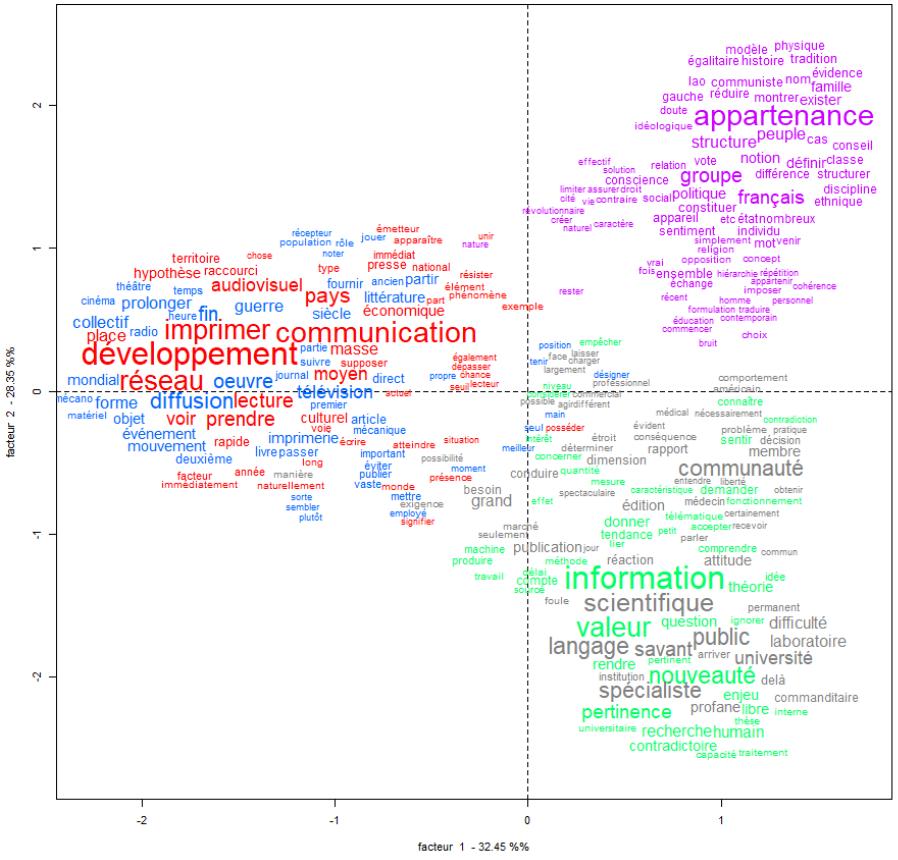

Figure 24: chi2-85

la digitalisation de l'oeuvre à son maximum, ce qui, pour les **Humanités Digitales**, est essentiel. Avoir travaillé non plus sur un texte, mais sur un corpus entier montre la multitude de possibilités pour analyser, étudier, comparer des textes. Nous aurions pu aussi comparer avec d'autres méta-données, mais le temps manquant, nous avons choisi de rester sur les deux plus importantes à nos yeux. Merci d'avoir pris le temps de lire ce compte-rendu. Pour finir, nous vous donnons une citation de Robert Escarpit, que l'on peut retrouver dans sa *Lettre ouverte au Diable* :

> "Ne pas mentir, c'est dire ce qu'on sait, non ce qu'on croit savoir."

Mélanie AUBRY et Cécile PORTAL